View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

ACADEMIC PRESS

Available at WWW.MATHEMATICSWEB.ORG POWERED BY SCIENCE DIRECT®

Journal of Multivariate Analysis 84 (2003) 40–60

Journal of Multivariate Analysis

http://www.elsevier.com/locate/jmva

Optimal prediction for linear regression with infinitely many parameters

Alexander Goldenshluger[a] and Alexandre Tsybakov[b,*]

[a] *Department of Statistics, University of Haifa, Haifa 31905, Israel*

[b] *Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VI, BP 188, 4 place Jussieu, Paris 75252, Cedex 05, France*

Received 22 November 1999

Abstract

The problem of optimal prediction in the stochastic linear regression model with infinitely many parameters is considered. We suggest a prediction method that outperforms asymptotically the ordinary least squares predictor. Moreover, if the random errors are Gaussian, the method is asymptotically minimax over ellipsoids in ℓ_2. The method is based on a regularized least squares estimator with weights of the Pinsker filter. We also consider the case of dynamic linear regression, which is important in the context of transfer function modeling.
© 2003 Elsevier Science (USA). All rights reserved.

AMS 1991 subject classifications: 62G05; 62G20

Keywords: Linear regression with infinitely many parameters; Optimal prediction; Exact asymptotics of minimax risk; Pinsker filter

1. Introduction

Consider the regression model

$$y = \sum_{k=1}^{\infty} \beta_k x_k + \epsilon, \tag{1}$$

where $\{x_k\}_{k=1,2,\ldots}$ is a sequence of possible explanatory variables, y is the corresponding response, ϵ is the error, and $\beta = (\beta_1, \beta_2, \ldots) \in \ell_2$ is an unknown

*Corresponding author.

E-mail addresses: goldensh@stat.haifa.ac.il (A. Goldenshluger), tsybakov@ccr.jussieu.fr (A. Tsybakov).

0047-259X/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0047-259X(02)00006-4

regression sequence. Assume that $\{x_k\}$ and ϵ are random variables, and $E\epsilon = 0$ and $E\epsilon^2 = \sigma^2$; the stochastic series in (1) and later are assumed to converge in the mean squared sense. Suppose we are given n realizations of y and $\{x_k\}$,

$$\{y(t); x_1(t), x_2(t), \ldots; t = 1, \ldots, n\} \tag{2}$$

coming from model (1); that is,

$$y(t) = \sum_{k=1}^{\infty} \beta_k x_k(t) + \epsilon(t), \quad t = 1, \ldots, n,$$

where $\epsilon(t), t = 1, 2, \ldots$, are i.i.d. random copies of ϵ and for each k the variables $x_k(t)$, $t = 1, 2, \ldots$, have the same distribution as x_k. Given $x_1(n+1), x_2(n+1), \ldots$, the objective is to predict the corresponding response $y(n+1)$ using data (2).

Following Breiman and Friedman [2], we establish our main results under the assumption that model (1) is canonical, i.e., $\{x_k\}$ are uncorrelated zero mean variables with variance 1. This assumption is not so restrictive in the prediction context. If $\{x_k\}$ are correlated then the standard Gram–Schmidt orthonormalizing process can be applied to get a canonical model with some other coefficient sequence (cf. [2]). Under the canonical formulation the influence of a particular regressor x_k on y is quantified solely by the magnitude of the corresponding coefficient β_k. We assume that the coefficients β_k are small for large values of k. Depending on the prior assumptions on the sequence β, only a certain finite number of first coefficients β_k is significant and should be kept for prediction. In Section 3 we indicate how the main results can be extended to the case of correlated regressors.

A *prediction method* (or *predictor*) $\hat{y}(n+1)$ is, in general, a random variable measurable with respect to $(\mathscr{U}_n, \mathscr{X}_{n+1})$ where $\mathscr{U}_n = \{y(t); x_1(t), x_2(t), \ldots; t = 1, \ldots, n\}$ and $\mathscr{X}_{n+1} = \{x_1(n+1), x_2(n+1), \ldots\}$. An important subclass of predictors that we call *natural predictors* and denote $\hat{y}^N(n+1)$ is defined by

$$\hat{y}^N(n+1) = \sum_{k=1}^{\infty} \hat{\beta}_k x_k(n+1), \tag{3}$$

where $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots)$ is an estimate for the regression coefficients $\beta = (\beta_1, \beta_2, \ldots)$. If $\hat{\beta}$ is a linear estimate, $\hat{y}^N(n+1)$ is called a *linear predictor*. The use of predictor (3) in practice is possible if only a finite number of (first) estimates $\hat{\beta}_k$ are non-zero. This is the case for the ordinary least squares (OLS) predictor, where $\hat{\beta}_k$ are the least squares estimators of β_k for $k \leqslant p$ and $\hat{\beta}_k = 0$ for $k > p$, with some given p [2,20].

In this paper we are interested in the optimal choice of predictor $\hat{y}$ in a minimax sense on a given family $\mathscr{B}$ of regression sequences β. The prediction error of $\hat{y}$ is defined as usually in the form $E[\hat{y}(n+1) - y(n+1)]^2$. Note that this error cannot be arbitrarily small; it is at least σ^2 for large n, because of the non–vanishing innovation component $\epsilon(n+1)$ independent of $(\mathscr{U}_n, \mathscr{X}_{n+1})$. We therefore consider the difference $E[\hat{y}(n+1) - y(n+1)]^2 - \sigma^2$, and define the maximal risk over $\mathscr{B}$ in

the form

$$\mathscr{R}[\hat{y};\mathscr{B}] = \sup_{\beta\in\mathscr{B}} E[\hat{y}(n+1) - y(n+1)]^2 - \sigma^2. \tag{4}$$

The optimal (minimax) predictor $\hat{y}_* = \hat{y}_*(n+1)$ minimizes the maximal prediction error,

$$\mathscr{R}[\hat{y}_*;\mathscr{B}] = \mathscr{R}_*[n;\mathscr{B}] \equiv \inf_{\hat{y}} \mathscr{R}[\hat{y};\mathscr{B}],$$

where inf is taken over all possible prediction methods based on the observations $(\mathscr{U}_n, \mathscr{X}_{n+1})$. Our aim is to find an *asymptotically minimax* prediction method $\hat{y}$ satisfying

$$\mathscr{R}[\hat{y};\mathscr{B}] = \mathscr{R}_*[n;\mathscr{B}](1+o(1)), \quad n\to\infty.$$

We will assume that $\mathscr{B}$ is an ellipsoid in the sequence space ℓ_2,

$$\mathscr{B} = \mathscr{B}(\{a_k\}, L) = \left\{\beta\in\ell_2 : \sum_{k=1}^{\infty} a_k^2\beta_k^2 \leqslant L^2\right\},$$

where $\{a_k\}_{k=1,2,\ldots}$ are positive coefficients such that $\{a_k\}$ is monotone non-decreasing, and $a_k\to\infty$ as $k\to\infty$. This assumption is natural since β_k, $k = 1,2,\ldots$, are the coefficients of the canonical regression model.

The main result of this paper consists in a construction of asymptotically minimax prediction methods (AMPM) for ellipsoids. We show that the AMPM is based not on the least squares estimator of β, but on properly weighted least squares, with the weights defined by the filter of Pinsker [17]. The AMPM outperforms the ordinary least squares predictor. The lower bound is proved for the case of Gaussian noise ϵ. It should be noticed also that construction of the proposed AMPM uses a priori information on the ellipsoid $\mathscr{B}$. In a subsequent paper [8], we develop an adaptive AMPM that does not require any a priori information on the class of sequences and is asymptotically sharp minimax on any ellipsoid within a wide scale. The idea behind construction of the adaptive AMPM is to apply the blockwise Stein rule to a penalized least squares estimate of the regression sequence. The corresponding block sizes increase "weakly" geometrically in order to ensure asymptotic minimaxity.

Our result is related to previous work in two aspects. First, the regression models with growing or infinite number of parameters have been studied by several authors [2,13,18–20,22]. In particular, Shibata [20], and Breiman and Friedman [2] develop the methods of optimal selection of the number of terms in a finite approximation to (1), using the ordinary least squares prediction. Shibata [20] considers the deterministic explanatory variables, while Breiman and Friedman [2] study the case where both $\{x_k\}_{k=1,2\ldots}$ and ϵ are Gaussian. Huber [13], Yohai and Maronna [22], and Portnoy [18,19] analyzed somewhat different setup. They consider regression with a finite but growing number of parameters p and obtain asymptotics for OLS

and, more generally, M-estimators in this model. Unlike our approach, this literature does not study weighted least squares prediction.

Second, the Pinsker filter has been extensively studied for different models, such as non-parametric regression and density estimation [1,4–6,12,15,21]. Golubev and Pinsker [10,11] develop asymptotically minimax methods for prediction of deterministic sequences observed without noise and with Gaussian white noise, respectively. Among this literature, the paper of Efromovich [4] that treats non-parametric regression with random design is closest to our setup. His paper considers the estimation of the vector of coefficients β from the observations $y(t) = \sum_{k=1}^{\infty} \beta_k \varphi_k(x(t)) + \epsilon(t),\ t = 1, \ldots, n,$ where $\{\varphi_k(\cdot)\}_{k=1,2,\ldots}$ is the orthonormal trigonometric basis on $[0, 1]$, and $x(t)$ are independent random variables distributed on $[0, 1]$. If their distribution is uniform, this is a special case of our model. On the other hand, we study prediction rather than estimation, and our method is different from that of Efromovich [4]. In particular, we do not use a two-stage procedure with preliminary consistent estimates. Furthermore, we discuss the problem with dependent observations, namely the dynamic linear regression where the response y is obtained as a convolution of the regression sequence β with the time sequence of explanatory variables. Such models arise in time-series analysis, linear system identification, and other applications (see, e.g., [3, Chapter 13; 14]).

2. Main results

In this section we assume that $x_k(t),\ t = 1, 2, \ldots$ are i.i.d. random variables for each k. Moreover, the following assumptions will be imposed on the explanatory variables and the errors of the model.

Assumption 1. The random variables $\{x_k\}_{k=1,2,\ldots}$ are uncorrelated, $Ex_k = 0,\ Ex_k^2 = 1$, and either

(i) $|x_k| \leqslant \kappa < \infty,\ \forall k$, or
(ii) $E|x_k|^{2p} \leqslant c^{2p-2}(2p)!$, for some $c > 0,\ p = 2, 3, \ldots,$ and $\forall k$.

Assumption 2. $E\epsilon = 0,\ E\epsilon^2 = \sigma^2,\ E\epsilon^4 \leqslant \tau\sigma^4 < \infty$ for some positive τ, and ϵ is independent of $\{x_k\}_{k=1,2,\ldots}$.

In order to define our prediction method we need the following notation (cf. [17]). Let v_n denote the solution of the equation

$$\sigma^2 n^{-1} \sum_{k=1}^{\infty} a_k (1 - v_n a_k)_+ = v_n L^2 \tag{5}$$

(note that the solution is unique since a_k are non-decreasing and $a_k \to \infty$). Let

$$\lambda_k = (1 - v_n a_k)_+, \quad k = 1, 2, \ldots, \tag{6}$$

$d_n(\mathscr{B}) \equiv \max\{k : a_k \leqslant v_n^{-1}\}$, and $\Lambda_d = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. In what follows, for brevity, we will write d or d_n instead of $d_n(\mathscr{B})$, keeping in mind that d depends both on the sample size n and on the class of regression sequences $\mathscr{B}$.

Denote $\phi_d(t) = (x_1(t), \ldots, x_d(t))'$, $t = 1, \ldots, n$, and let

$$\tilde{b}_d = \left(\frac{1}{n}\sum_{t=1}^{n} \phi_d(t)\phi_d'(t) + n^{-1}I_d\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} \phi_d(t)y(t)\right)$$
$$\equiv Q_d^{-1}(n)\left(\frac{1}{n}\sum_{t=1}^{n} \phi_d(t)y(t)\right), \tag{7}$$

where I_d stands for the identity $d \times d$ matrix. In fact, $\tilde{b}_d = (\tilde{\beta}_1, \ldots, \tilde{\beta}_d)'$ is a regularized version of the standard least squares estimate for the vector $b_d = (\beta_1, \ldots, \beta_d)'$ composed of the first d coefficients of the regression sequence β. We introduce this regularization in order to improve the behavior of $\tilde{b}_d$ for small sample sizes n, when the matrix $n^{-1}\sum_{t=1}^{n} \phi_d(t)\phi_d'(t)$ may not be well conditioned.

Define

$$\hat{\beta}_* = (\hat{\beta}_1, \ldots, \hat{\beta}_d; 0, 0, \ldots) \equiv (\tilde{b}_d'\Lambda_d; 0, 0, \ldots). \tag{8}$$

Let $\hat{y}_* = \hat{y}_*(n+1)$ be the predictor given by (3) with $\hat{\beta} = \hat{\beta}_*$ as in (8). Note that $\hat{y}_*$ is a linear predictor with finite number d of summands in (3), and it is different from the OLS predictor. The predictor $\hat{y}_*$ is optimal in the following minimax sense.

Theorem 1. *Let Assumptions* 1 *and* 2 *hold, and*

$$d_n\sqrt{\ln(n)/n} \to 0, \quad n \to \infty. \tag{9}$$

Assume also that $k^{-1/2}a_k \to \infty$ *as* $n \to \infty$. *Then*

$$\mathscr{R}[\hat{y}_*; \mathscr{B}] \leqslant r_n(1 + o(1)), \quad n \to \infty, \tag{10}$$

where

$$r_n = r_n(\mathscr{B}) = \sigma^2 n^{-1}\sum_{k=1}^{\infty} (1 - v_n a_k)_+.$$

Consider an application of Theorem 1. Let $a_k = k^\alpha$, $k = 1, 2, \ldots$, $\alpha > 1/2$, and let $\mathscr{B}_\alpha$ denote the ellipsoid $\mathscr{B}(\{k^\alpha\}, L)$. Then

$$v_n = \left[\frac{(\alpha+1)(2\alpha+1)}{\alpha} L^2\sigma^{-2}n\right]^{-\frac{\alpha}{2\alpha+1}}(1 + o(1)),$$

and $d_n = O(n^{1/(2\alpha+1)})$ as $n \to \infty$. Thus, (9) is satisfied. The asymptotical maximal risk of our prediction method $\hat{y}_*$ is

$$r_n(\mathscr{B}_\alpha) = C^*(\alpha)L^{\frac{2}{2\alpha+1}}\left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}(1 + o(1)),$$

where

$$C^*(\alpha) = \left(\frac{\alpha}{\alpha+1}\right)^{\frac{2\alpha}{2\alpha+1}} (2\alpha+1)^{\frac{1}{2\alpha+1}}$$

is the Pinsker constant. Let us compare now this risk to the maximal risk of the ordinary least squares (OLS) predictor. Confining ourselves to the Gaussian case and using the results of Breiman and Friedman [2], we find that the error of the OLS predictor of the order p (denoted $\hat{y}_p^{\mathrm{OLS}}$) is

$$E[\hat{y}_p^{\mathrm{OLS}}(n+1) - y(n+1)]^2 - \sigma^2 = \left(\sum_{k>p} \beta_k^2 + \frac{p\sigma^2}{n}\right)(1+o(1)), \quad n\to\infty.$$

Thus,

$$\mathscr{R}[\hat{y}_p^{\mathrm{OLS}}; \mathscr{B}_\alpha] = \left(L^2 p^{-2\alpha} + \frac{p\sigma^2}{n}\right)(1+o(1)).$$

The maximal risk of the best OLS predictor is

$$\min_p \mathscr{R}[\hat{y}_p^{\mathrm{OLS}}; \mathscr{B}_\alpha] = \left(\frac{2\alpha+1}{2\alpha}\right)(2\alpha)^{\frac{1}{2\alpha+1}} L^{\frac{2}{2\alpha+1}} \left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}} (1+o(1)),$$

and it is easy to see that this expression is greater than $r_n(\mathscr{B}_\alpha)$:

$$\frac{\min_p \mathscr{R}[\hat{y}_p^{\mathrm{OLS}}; \mathscr{B}_\alpha]}{r_n(\mathscr{B}_\alpha)} = \left[\frac{(2\alpha+1)(\alpha+1)}{2\alpha^2}\right]^{\frac{2\alpha}{2\alpha+1}} (1+o(1)) > 1.$$

The result of Theorem 1 cannot be improved among all prediction methods in the case where ϵ is Gaussian. We now state the lower bound showing this fact.

Assumption 3. The random variable ϵ is Gaussian $\mathscr{N}(0,\sigma^2)$, and ϵ is independent of $\{x_k\}_{k=1,2,\ldots}$.

Theorem 2. *Let Assumptions* 1 *and* 3 *hold. Assume that either*

$$\exp\left\{-\frac{\gamma n^2 v_n^2}{\sum_{k=1}^{d_n} a_k^2 (1 - v_n a_k)_+^2}\right\} = o(v_n), \quad n\to\infty, \ \forall \gamma > 0 \tag{11}$$

or

$$v_n \sum_{k=1}^{d_n} a_k = o(d_n), \quad n\to\infty. \tag{12}$$

Then for every prediction method $\hat{y} = \hat{y}(n+1)$ *one has*

$$\mathscr{R}[\hat{y}; \mathscr{B}] \geqslant r_n(1+o(1)), \quad n\to\infty. \tag{13}$$

It can be easily verified that (11) is valid for the ellipsoids with polynomially increasing sequences $\{a_k\}$, while (12) holds for exponentially increasing $\{a_k\}$.

Thus, Theorem 1 along with Theorem 2 shows that predictor (3) associated with $\hat{\beta}_*$ given by (8) is asymptotically minimax.

3. Correlated regressors

In this section we indicate how the above results can be extended to the case of correlated regressors. We also consider an important specific example of dynamic linear regression model where the time sequence of explanatory variables is correlated.

3.1. Reduction to a canonical model

Consider the regression model

$$y = \sum_{k=1}^{\infty} \theta_k z_k + \varepsilon,$$

where $\{z_k\}_{k=1,2,\dots}$ is a sequence of explanatory variables and $\theta = (\theta_1, \theta_2, \dots) \in \ell_2$ is an unknown regression sequence. As before, given $\{y(t); z_1(t), z_2(t), \dots; t = 1, \dots, n\}$ and $z_1(n+1), z_2(n+1), \dots$, we wish to predict $y(n+1)$. In contrast to the canonical model (1) we assume here that the regressors $\{z_k\}_{k=1,2,\dots}$ are correlated. Since we are interested in prediction, we can represent the random variable $\xi = \sum_{k=1}^{\infty} \theta_k z_k$ in an orthonormal basis, passing thus to a canonical model. Observe that if θ belongs to an ellipsoid $\mathscr{B}$, then the coefficient sequence of the corresponding canonical model does not necessarily belongs to $\mathscr{B}$. Nevertheless, under mild conditions on the correlation between regressors, the ellipsoidal structure of the problem is preserved when passing to the canonical model.

Let $z_1, z_2, \dots$ be correlated, $Ez_k = 0$, $Ez_k^2 = 1$, and any finite number of elements $z_1, z_2, \dots, z_k$ be linearly independent. Assume that $\theta \in \mathscr{B}(\{c_k\}, L)$ with monotone increasing sequence $\{c_k\}$ satisfying $c_k k^{-2\gamma} \to \infty$ as $k \to \infty$ for some $\gamma > 1/2$. Then the coefficient sequence β of the corresponding canonical model belongs to any ellipsoid $\mathscr{B}(\{a_k\}, LM)$ such that $\sum_{k=1}^{\infty} a_k^2 c_k^{-2} \leqslant M^2 < \infty$. Indeed, the standard Gram–Schmidt orthonormalizing process yields the orthonormal basis $\{x_k\}$ with the following properties. There exist constants $\{h_{jk}\}_{j,k=1,2,\dots}$, such that

$$z_j = \sum_{k=1}^{j} h_{jk} x_k, \quad j = 1, 2, \dots, \tag{14}$$

and $Ex_k = 0$, $Ex_k x_j = \delta_{jk}$, where δ_{jk} stands for the Kronecker symbol. It follows from the construction that $h_{jk} = E(z_j x_k)$ and $h_{jk} = 0$, for $k > j$. The random variable $\xi = \sum_{k=1}^{\infty} \theta_k z_k$ is represented in the orthonormal basis $(x_1, x_2, \dots)$ as $\xi = \sum_{k=1}^{\infty} E(\xi x_k) x_k = \sum_{k=1}^{\infty} \beta_k x_k$, and the regression sequence $\beta = (\beta_1, \beta_2, \dots)$ in the

canonical model is given by

$$\beta_k = E(\xi x_k) = \sum_{j=1}^{\infty} \theta_j h_{jk} = \sum_{j=k}^{\infty} \theta_j h_{jk}.$$

Further, note that the condition $Ez_k^2 = 1$ amounts to $\sum_{k=1}^{j} h_{jk}^2 = 1, \ \forall k$ (see (14)). Let $\{a_k\}$ be a positive monotone increasing sequence such that $\sum_{k=1}^{\infty} a_k^2 c_k^{-2} \leqslant M^2 < \infty$. For example, if $c_k k^{-2\gamma} \to \infty$ as $k \to \infty$ for some $\gamma > 1/2$, then $a_k = c_k k^{-\gamma}$ can be taken. By the Cauchy–Schwarz inequality and monotonicity of $\{a_k\}$

$$\sum_{k=1}^{\infty} a_k^2 \beta_k^2 \leqslant L^2 \sum_{k=1}^{\infty} a_k^2 \sum_{j=k}^{\infty} h_{jk}^2 c_j^{-2} = L^2 \sum_{j=1}^{\infty} \sum_{k=1}^{j} a_k^2 h_{jk}^2 c_j^{-2}$$

$$\leqslant L^2 \sum_{j=1}^{\infty} a_j^2 c_j^{-2} \sum_{k=1}^{j} h_{jk}^2 = L^2 \sum_{j=1}^{\infty} a_j^2 c_j^{-2} \leqslant L^2 M^2.$$

Thus $\beta \in \mathscr{B}(\{a_k\}, LM)$ as claimed.

Although we have the above relationship between the coefficient sequences, it is not unreasonable to impose ellipsoidal constraints directly on the coefficients β of the canonical model. In fact, the influence of the corresponding regressor x_k on the response y is quantified solely by the magnitude of β_k. In this case, the results about the statistical properties of our prediction method are exactly the same as in Section 2.

Note that the above reduction to the canonical model applies only when correlations between the original regressors $\{z_k\}$ are known. Otherwise a sampled version of the orthonormalizing process can be performed in a standard way. Observe that it is sufficient to "decorrelate" the $d = d_n$ first regressors, because the AMPM is based only on the d first "principle components". In this case the corresponding prediction method can be defined similarly to (3), (6)–(8) with the following modifications. Let $\phi_d(t) = (z_1(t), \ldots, z_d(t))', \ t = 1, \ldots, n$,

$$\hat{\Sigma}_d = \left(\frac{1}{n} \sum_{t=1}^{n} \phi_d(t) \phi_d(t)' + n^{-1} I_d \right), \quad \tilde{b}_d = \hat{\Sigma}_d^{-1/2} \left(\frac{1}{n} \sum_{t=1}^{n} \phi_d(t) y(t) \right).$$

Let $\hat{\beta}_*$ be given by (8) and $\tilde{\phi}_d(t) = (x_1(t), \ldots, x_d(t))' = \hat{\Sigma}_d^{-1/2} \phi_d(t)$. Then the predictor is defined by $\hat{y}(n+1) = \hat{\beta}_*' \tilde{\phi}_d(n+1)$. We conjecture that this prediction method is asymptotically minimax in the case of correlated regressors.

3.2. Dynamic linear regression model

In many applications the following dynamic linear regression model is of interest:

$$y(t) = \sum_{k=1}^{\infty} \beta_k u(t-k) + \epsilon(t), \quad t = 1, \ldots, n. \tag{15}$$

In the context of time-series analysis one can think of (15) as being the transfer function model between two time series $\{y(t)\}$ and $\{u(t)\}$ (cf. [3, Section 13.1]). For example, model (15) contains as a special case (but is not limited to) the state space model $y(t) = u(t) + \epsilon(t)$ with an ARMA (p, q) process $u(t)$. Of course, in this case the coefficients β_k should be exponentially decreasing. Polynomially decreasing β_k (allowed by our model) correspond to long-range dependence.

Minimax rates of convergence in estimating $\beta = (\beta_1, \beta_2, \dots)$ under model (15) have been studied recently by Goldenshluger [7]. Here we consider the prediction problem and propose a different method that achieves not only the rates but also the exact minimax constants.

Given the data $\mathscr{U}_n = \{y(t), u(t-1); t = 2, \dots, n\}$ our objective is to predict the output (response) $y(n+1)$. A predictor $\hat{y}(n+1)$ can be any random variable measurable w.r.t. $(\mathscr{U}_n, u(n))$. In contrast to (1), the vectors of the explanatory variables in (15) are dependent. It turns out that the results of Section 2 can be extended for the case of the dynamic linear regression model.

We use the same notation as in Section 2; the only difference is that now

$$\phi_d(t) = (u(t-1), \dots, u(t-d))', \quad t = 1, \dots, n, \tag{16}$$

and that the vectors $\phi_d(t)$ can involve inputs $u(t)$ for $t \leqslant 0$: in this case the inputs are assumed to be replaced by zeros in (16). Define the prediction method $\hat{y}_*$ by the same formulae as in Section 2. As before, the maximal risk $\mathscr{R}[\hat{y}; \mathscr{B}]$ is given by (4).

Assumption 1′. The random variables $u(t)$, $t = \dots, -1, 0, 1, \dots$, are independent and identically distributed, $Eu(t) = 0$, $E|u(t)|^2 = 1$, and either

(i) $|u(t)| \leqslant \kappa < \infty, \forall t$, or
(ii) $E|u(t)|^{2p} \leqslant c^{2p-2}(2p)!$, for some $c > 0$, $p = 2, 3, \dots$.

Assumption 2′. The random variables $\epsilon(t)$, $t = 1, 2, \dots$, are independent identically distributed, independent of $\{u(t)\}$, and $E\epsilon(t) = 0$, $E|\epsilon(t)|^2 = \sigma^2$, $E|\epsilon(t)|^4 \leqslant \tau\sigma^4 < \infty$ for some positive τ.

The next theorem is an analog of Theorem 1 for the dynamic regression model.

Theorem 3. *Let Assumptions 1′ and 2′ hold, and* $d_n\sqrt{\ln(n)/n} \to 0$ *as* $n \to \infty$. *Assume that* $k^{-1/2}a_k \to \infty$ *as* $k \to \infty$. *Then*

$$\mathscr{R}[\hat{y}_*; \mathscr{B}] \leqslant r_n(1 + o(1)), \quad n \to \infty.$$

Remark. Goldenshluger and Zeevi [9] study minimax rates of prediction for autoregressive models with infinitely many parameters β_k. Their setup is different from the regression setup (15) and, furthermore, it is restricted to exponentially decreasing β_k. Note also that the method of Goldenshluger and Zeevi [9] does not

involve the Pinsker filter and, unlike Theorem 3, their result concentrates on non-asymptotic bounds and does not give the asymptotically exact constants.

4. Numerical results

A small simulation study has been conducted to illustrate the practical behavior of the proposed asymptotically minimax prediction method (AMPM). It is expected that for a given ellipsoid the AMPM will outperform the best ordinary least squares (OLS) predictor when the regression sequence β is close to the worst–case sequence from the class. The goal of the following is to understand for which sample sizes and ellipsoids the difference between the methods becomes apparent.

In the simulation study we consider the ellipsoids $\mathscr{B}(\{a_k\}, L)$ with $a_k = k^\alpha$, $k = 1, 2, \ldots$ and $L = 1$. The data $(\mathscr{U}_n, \mathscr{X}_{n+1})$ are generated from the canonical model (1), where $\varepsilon \sim \mathscr{N}(0,1)$ and the regressors $\{x_k\}$ are i.i.d. $\mathscr{N}(0,1)$ random variables. The regression sequence β is chosen in the following way: β_k are independent Gaussian random variables with zero mean and variance $m_k^2 = (1 - v_n k^\alpha)_+ (n v_n k^\alpha)^{-1}$, where v_n is given by (5) with $a_k = k^\alpha$ and $L = 1$. If $m_k^2 = 0$ then we set $\beta_k = 0$. Such a sequence with large probability belongs to the ellipsoid $\mathscr{B}(\{k^\alpha\}, 1)$ (see proof of Theorem 2 below). For given α and n we compute the root of the mean squared prediction risk averaged over $N = 2000$ replications. Recall that in our case the mean squared prediction risk of a method $\hat{y} = \hat{y}(n+1)$ is defined by

$$\mathscr{R}[\hat{y}, y] = E[\hat{y}(n+1) - y(n+1)]^2 - 1.$$

The results for the AMPM and the best OLS predictor appear in Table 1. We display the values for $\alpha = 1$ and 2. Simulation shows that, as expected, the AMPM

Table 1
The root of the mean squared prediction risk for 2000 replications

	n	AMPM	OLS
$\alpha = 1$			
	50	0.229	0.303
	75	0.198	0.276
	100	0.180	0.245
	200	0.156	0.201
	500	0.126	0.144
	1000	0.092	0.117
$\alpha = 2$			
	50	0.213	0.265
	75	0.188	0.209
	100	0.160	0.179
	200	0.128	0.142
	500	0.092	0.135
	1000	0.068	0.086

outperforms the best OLS. This is apparent even for comparatively small sample sizes. We observed that the difference in performance is especially pronounced for small values of α, i.e., for more heavy-tailed sequences β.

5. Proofs

5.1. Proof of Theorem 1

We give the proof under Assumption 1(i) only. Under Assumption 1(ii) the proof is essentially the same; only minor modifications should be made. First, in the proof of Lemma 1 below one needs to use the Bernstein exponential inequality instead of the Hoeffding one (see [16, Chapter 2]). Second, Lemmas 2 and 3 hold true with some new constants depending on the moment growth conditions for $\{x_k\}$. The corresponding bounds are easily obtained using the Cauchy–Schwarz inequality.

1. By Assumptions 1 and 2

$$E[\hat{y}_*(n+1) - y(n+1)]^2 = E\left[\sum_{k=1}^{\infty} (\hat{\beta}_k - \beta_k)x_k(n+1) + \epsilon(n+1)\right]^2$$
$$= E||\hat{\beta} - \beta||_2^2 + \sigma^2,$$

where $||\cdot||_2$ denotes the standard norm in the sequence space ℓ_2. Therefore it is sufficient to bound from above $\sup_{\beta\in\mathscr{B}} E||\hat{\beta} - \beta||_2^2$. First, we note that

$$E||\hat{\beta} - \beta||_2^2 = E||\Lambda_d \tilde{b}_d - b_d||_2^2 + \sum_{k=d+1}^{\infty} \beta_k^2 \tag{17}$$

and

$$\tilde{b}_d - b_d = Q_d^{-1}(n)\left(-n^{-1}b_d + \frac{1}{n}\sum_{t=1}^{n} \phi_d(t) \sum_{k=d+1}^{\infty} \beta_k x_k(t) + \frac{1}{n}\sum_{t=1}^{n} \phi_d(t)\epsilon(t)\right). \tag{18}$$

Further,

$$E||\Lambda_d \tilde{b}_d - b_d||_2^2$$
$$= b_d'(I_d - \Lambda_d)^2 b_d + E[(\tilde{b}_d - b_d)'\Lambda_d^2(\tilde{b}_d - b_d)] - 2b_d'\Lambda_d(I_d - \Lambda_d)E(\tilde{b}_d - b_d)$$
$$= \sum_{k=1}^{d} (1-\lambda_k)^2\beta_k^2 + \sum_{k=1}^{d} \lambda_k^2 E(\tilde{\beta}_k - \beta_k)^2 - 2\sum_{k=1}^{d} \lambda_k(1-\lambda_k)\beta_k E(\tilde{\beta}_k - \beta_k)$$
$$\equiv \sum_{k=1}^{d} (1-\lambda_k)^2\beta_k^2 + I_1(n,\beta) + I_2(n,\beta) \tag{19}$$

(recall that $\tilde{b}_d = (\tilde{\beta}_1, \ldots, \tilde{\beta}_d)'$).

2. Let $||\cdot||_\Lambda$ denote the norm on the space of real–valued sequences ℓ_2 which is generated by the inner product $\langle x, y\rangle_\Lambda = \sum_{k=1}^{\infty} \lambda_k^2 x_k y_k$; here $\{\lambda_k\}_{k=1,2,\ldots}$ are defined

in (6). In fact, $||\cdot||_\Lambda$ is a semi-norm on ℓ_2, but an actual norm on the d-dimensional linear subspace of ℓ_2. This vector norm defines the corresponding operator matrix norm, and our current goal is to establish useful bounds on $||Q_d^{-1}(n)||_\Lambda$.

Let $v_{ij} \equiv n^{-1}\sum_{t=1}^n x_i(t)x_j(t) - \delta_{ij}$, where $i,j = 1,\dots,d$, and δ_{ij} stands for the Kronecker symbol. Fix $\alpha \in (0,1)$ and define the random event

$$\Omega_\alpha \equiv \left\{\omega \in \Omega : \max_{i,j=1,\dots,d} |v_{ij}| \leqslant m_n(\alpha)\right\}, \quad m_n(\alpha) = \frac{\kappa^2}{\sqrt{2n}}\sqrt{\ln\frac{2d^2}{\alpha}}. \tag{20}$$

We have the following lemma.

Lemma 1. *Let Assumptions* 1 *and* 2 *hold*; *then*

$$P(\Omega_\alpha) \geqslant 1-\alpha. \tag{21}$$

If

$$\rho_n(\alpha) \equiv \frac{d_n}{n} + d_n m_n(\alpha) < 1, \tag{22}$$

then on the event Ω_α

$$1 - \frac{\rho_n(\alpha)}{1+\rho_n(\alpha)} \leqslant ||Q_d^{-1}(n)||_\Lambda \leqslant 1 + \frac{\rho_n(\alpha)}{1-\rho_n(\alpha)}. \tag{23}$$

On the complementary event $\bar{\Omega}_\alpha$, $||Q_d^{-1}(n)||_\Lambda \leqslant n$.

Proof. First we prove (21). For any $\epsilon > 0$ we have

$$P\left\{\max_{i,j=1,\dots,d} |v_{ij}| \geqslant \epsilon\right\} \leqslant P\left\{\max_{i=1,\dots,d}\left|\frac{1}{n}\sum_{t=1}^n x_i^2(t) - 1\right| \geqslant \epsilon\right\} + P\left\{\max_{i,j=1,\dots,d, i\neq j}\left|\frac{1}{n}\sum_{t=1}^n x_i(t)x_j(t)\right| \geqslant \epsilon\right\} \equiv P_1 + P_2.$$

Now observe that $\{n^{-1}(x_i^2(t)-1)\}_{t=1,\dots,n}$ is a sequence of i.i.d zero mean random variables with bounded ranges. By Hoeffding's inequality

$$P\left\{\left|\frac{1}{n}\sum_{t=1}^n x_i^2(t) - 1\right| \geqslant \epsilon\right\} \leqslant 2\exp\left(-\frac{2\epsilon^2 n}{\kappa^4}\right)$$

and therefore $P_1 \leqslant 2d\exp(-2\epsilon^2 n/\kappa^4)$. Similarly we have $P_2 \leqslant 2d(d-1)\exp(-2\epsilon^2 n/\kappa^4)$, and finally

$$P\left\{\max_{i,j=1,\dots,d} |v_{ij}| \geqslant \epsilon\right\} \leqslant 2d^2\exp\left(-\frac{2\epsilon^2 n}{\kappa^4}\right).$$

Setting $\epsilon = m_n(\alpha)$ we come to (21). Define

$$A_d(n) \equiv I_d - Q_d(n) = I_d - \frac{1}{n}\sum_{t=1}^n \phi_d(t)\phi_d'(t) - \frac{1}{n}I_d.$$

If the event Ω_α holds, then $|[A_d(n)]_{ij}| \leqslant n^{-1} + m_n(\alpha),\ i,j = 1, \ldots, d$ (here $[A_d(n)]_{ij}$ denotes the i,j-entry of the matrix A_d). Let $A_d^j(n),\ j = 1, \ldots, d$ denote the rows of the matrix $A_d(n)$. Then it is easily checked that $||A_d(n)||_\Lambda \leqslant (\sum_{j=1}^d ||A_d^j(n)||_2^2)^{1/2}$. Thus, on the set Ω_α we have

$$||A_d(n)||_\Lambda \leqslant \frac{d_n}{n} + d_n m_n(\alpha).$$

Due to (22), $||A_d(n)||_\Lambda < 1$ and therefore

$$\frac{1}{1 + ||A_d(n)||_\Lambda} \leqslant ||(I_d - A_d(n))^{-1}||_\Lambda = ||Q_d^{-1}(n)||_\Lambda \leqslant \frac{1}{1 - ||A_d(n)||_\Lambda}.$$

Using the above bound on $||A_d(n)||_\Lambda$ we obtain (23). The lemma is proved. □

3. The next step in our analysis is to bound from above the quantity $I_1(n,\beta) = E||\tilde{b_d} - b_d||_\Lambda^2$ (see (19)). First, we establish some useful bounds that will be used later. We have

$$||\tilde{b_d} - b_d||_\Lambda^2 \leqslant ||Q_d^{-1}(n)||_\Lambda^2|| - n^{-1}b_d + I_{11} + I_{12}||_\Lambda^2, \tag{24}$$

where

$$I_{11} \equiv \frac{1}{n}\sum_{t=1}^n \phi_d(t) \sum_{k=d+1}^\infty \beta_k x_k(t), \tag{25}$$

$$I_{12} \equiv \frac{1}{n}\sum_{t=1}^n \phi_d(t)\epsilon(t). \tag{26}$$

Lemma 2. *Under Assumptions* 1 *and* 2

$$||n^{-1}b_d||_\Lambda^2 = n^{-2}\sum_{k=1}^d \lambda_k^2\beta_k^2, \tag{27}$$

$$E||I_{11}||_\Lambda^2 \leqslant \frac{\kappa^2}{n}\sum_{k=1}^d \lambda_k^2 \sum_{j=d+1}^\infty \beta_j^2, \tag{28}$$

$$E||I_{12}||_\Lambda^2 = \frac{\sigma^2}{n}\sum_{k=1}^d \lambda_k^2. \tag{29}$$

In addition, $E\langle I_{11}, I_{12}\rangle_\Lambda = 0$.

Proof. The proof of (27) is straightforward. To show (28) we denote

$$I_{11,k} = \frac{1}{n}\sum_{t=1}^n x_k(t) \sum_{j=d+1}^\infty \beta_j x_j(t), \quad k = 1, \ldots, d.$$

Notice that $E[I_{11,k}]=0,\ \forall k=1,\dots,d$. We have by Assumptions 1 and 2

$$E[I_{11,k}]^2 = E\left(\frac{1}{n^2}\sum_{t,s=1}^{n} x_k(t)x_k(s)\sum_{j,l=d+1}^{\infty}\beta_j\beta_l x_j(t)x_l(s)\right)$$
$$=\frac{1}{n^2}E\sum_{t=1}^{n}x_k^2(t)\left(\sum_{j=d+1}^{\infty}\beta_j x_j(t)\right)^2 \leqslant \frac{\kappa^2}{n}\sum_{j=d+1}^{\infty}\beta_j^2,$$

and this implies (28).

Let $I_{12,k}=n^{-1}\sum_{t=1}^{n}x_k(t)\epsilon(t),\ k=1,\dots,d$; then $E||I_{12,k}||_{\Lambda}^2 = E\sum_{k=1}^{d}\lambda_k^2[I_{12,k}]^2$, and (29) follows by direct calculations. Notice also that $E[I_{12,k}]=0,\ \forall k=1,\dots,d$.

Further, for every $k=1,\dots,d$

$$E[I_{11,k}I_{12,k}] = E\left(\frac{1}{n^2}\sum_{t,s=1}^{n}x_k(s)\epsilon(s)x_k(t)\sum_{j=d+1}^{\infty}\beta_j x_j(t)\right)$$
$$=\frac{1}{n^2}\sum_{t=1}^{n}E\left(x_k^2(t)\epsilon(t)\sum_{j=d+1}^{\infty}\beta_j x_j(t)\right)=0$$

since ϵ is independent of $\{x_k\}_{k=1,2,\dots}$. Hence $E\langle I_{11},I_{12}\rangle_{\Lambda}=0$ as claimed. □

Lemma 3. *Suppose that Assumptions* 1 *and* 2 *hold, and* $\beta\in\ell_1$ *or* $\beta\in\mathscr{B}(\{a_k\},L)$ *with constants* a_k *satisfying* $k^{-1/2}a_k\to\infty,\ k\to\infty$. *Then there exist constants* C_1 *and* C_2 *depending on* $\sigma^2,\ \kappa,\ \tau$ *only such that*

$$(E||I_{11}||_{\Lambda}^4)^{1/2}\leqslant\frac{C_1}{n}\left(\sum_{k=d+1}^{\infty}|\beta_k|\right)^2\sum_{k=1}^{d}\lambda_k^2,\quad (E||I_{12}||_{\Lambda}^4)^{1/2}\leqslant\frac{C_2}{n}\sum_{k=1}^{d}\lambda_k^2.$$

Proof. We start with bounding $E||I_{11}||_{\Lambda}^4 = E(\sum_{k=1}^{d}\lambda_k^2[I_{11,k}]^2)^2$. For every $k=1,\dots,d$ we have, due to independence of the replications $\{x_1(t),x_2(t),\dots\}_{t=1,\dots,n}$,

$$E[I_{11,k}]^4 = E\left(\frac{1}{n}\sum_{t=1}^{n}x_k(t)\sum_{j=d+1}^{\infty}\beta_j x_j(t)\right)^4$$
$$=\frac{1}{n^4}E\left[\sum_{t,s=1}^{n}x_k^2(t)x_k^2(s)\left(\sum_{j=d+1}^{\infty}\beta_j x_j(t)\right)^2\left(\sum_{j=d+1}^{\infty}\beta_j x_j(s)\right)^2\right]$$
$$\leqslant\frac{\kappa^4}{n^4}\sum_{t,s=1}^{n}E\left[\left(\sum_{j=d+1}^{\infty}\beta_j x_j(t)\right)^2\left(\sum_{j=d+1}^{\infty}\beta_j x_j(s)\right)^2\right]$$
$$\leqslant\frac{\kappa^8}{n^2}\left(\sum_{j=d+1}^{\infty}|\beta_j|\right)^4.$$

Thus,

$$E||I_{11}||_{\Lambda}^4 = E\sum_{k,j=1}^{d} \lambda_k^2\lambda_j^2[I_{11,k}]^2[I_{11,j}]^2 \leqslant \frac{\kappa^8}{n^2}\left(\sum_{j=d+1}^{\infty} |\beta_j|\right)^4 \left(\sum_{k=1}^{d} \lambda_k^2\right)^2.$$

Similarly, for every $k = 1, \ldots, d$,

$$\begin{aligned} E[I_{12,k}]^4 &= E\left(\frac{1}{n}\sum_{t=1}^{n} x_j(t)\epsilon(t)\right)^4 = \frac{1}{n^4}\sum_{t,s=1}^{n} E[x_k^2(t)\epsilon^2(t)x_k^2(s)\epsilon^2(s)] \\ &= \frac{\sigma^4}{n^2}\left(1-\frac{1}{n}\right) + \frac{\kappa^4\sigma^4\tau}{n^3} \end{aligned}$$

and therefore

$$E||I_{12}||_{\Lambda}^4 = E\left(\sum_{k=1}^{d} \lambda_k^2[I_{12,k}]^2\right)^2 \leqslant \frac{\sigma^4}{n^2}\left(\sum_{k=1}^{d} \lambda_k^2\right)^2 \left(1+\frac{\kappa^4\tau}{n}\right).$$

This completes the proof. □

4. Now we are ready to establish an upper bound on $I_1(n,\beta) = E||\tilde{b}_d - b_d||_{\Lambda}^2$. Let the event Ω_α be defined by (20). We choose $\alpha = \alpha^* = 2d^2n^{-8}$, and let $\rho_n^* = \rho_n(\alpha^*)$. Note that condition (9) ensures (22) of Lemma 1 for large enough n. In addition, (9) implies $\rho_n^* \to 0$ as $n \to \infty$. It follows from (24) and Lemma 2 that

$$\begin{aligned} &E(||\tilde{b}_d - b_d||_{\Lambda}^2 \mathbf{1}\{\Omega_{\alpha^*}\}) \\ &\leqslant \left(1+\frac{\rho_n^*}{1-\rho_n^*}\right) E||-n^{-1}b_d + I_{11} + I_{12}||_{\Lambda}^2 \\ &= \left(1+\frac{\rho_n^*}{1-\rho_n^*}\right)(||n^{-1}b_d||_{\Lambda}^2 + E||I_{11}||_{\Lambda}^2 + E||I_{12}||_{\Lambda}^2) \\ &\leqslant \left(1+\frac{\rho_n^*}{1-\rho_n^*}\right)\left(\frac{1}{n^2}\sum_{k=1}^{d}\lambda_k^2\beta_k^2 + \frac{\kappa^2}{n}\sum_{k=1}^{d}\lambda_k^2\sum_{j=d+1}^{\infty}\beta_j^2 + \frac{\sigma^2}{n}\sum_{k=1}^{d}\lambda_k^2\right) \\ &\equiv J_1(n,\beta), \end{aligned} \tag{30}$$

where $\mathbf{1}\{\cdot\}$ stands for the indicator function.

Similarly, using Lemmas 1–3 and the Cauchy–Schwarz inequality,

$$\begin{aligned} E(||\tilde{b_d} - b_d||_\Lambda^2 \mathbf{1}\{\bar{\Omega}_{\alpha^*}\}) &\leqslant n^2 E(|| - n^{-1} b_d + I_{11} + I_{12}||_\Lambda^2 \mathbf{1}\{\bar{\Omega}_{\alpha^*}\}) \\ &\leqslant 4n^2 (||n^{-1} b_d||_\Lambda^2 P(\bar{\Omega}_{\alpha^*}) + E[||I_{11}||_\Lambda^2 \mathbf{1}\{\bar{\Omega}_{\alpha^*}\}] \\ &\quad + E[||I_{12}||_\Lambda^2 \mathbf{1}\{\bar{\Omega}_{\alpha^*}\}]) \\ &\leqslant 4n^2 \left[\frac{\alpha^*}{n^2} \sum_{k=1}^{d} \lambda_k^2 \beta_k^2 + \frac{C_3}{n} \sum_{k=1}^{d} \lambda_k^2 \left(\sum_{k=d+1}^{\infty} |\beta_k| \right)^2 \sqrt{\alpha^*} \right] \\ &\leqslant \frac{8d^2}{n^8} \sum_{k=1}^{d} \lambda_k^2 \beta_k^2 + \frac{C_3 d}{n^3} \sum_{k=1}^{d} \lambda_k^2 \left(\sum_{k=d+1}^{\infty} |\beta_k| \right)^2 \equiv J_2(n, \beta), \end{aligned} \tag{31}$$

where C_3 is a constant depending on σ^2, κ, and τ only. Thus, we obtain

$$I_1(n, \beta) = E||\tilde{b_d} - b_d||_\Lambda^2 \leqslant J_1(n, \beta) + J_2(n, \beta), \tag{32}$$

where $J_1(n, \beta)$ and $J_2(n, \beta)$ are given by (30) and (31), respectively.

5. Taking into account (30) and (31) and returning to (17) and (19) we can write

$$\begin{aligned} \sup_{\beta \in \mathscr{B}} E||\hat{\beta} - \beta||_2^2 &\leqslant \sup_{\beta \in \mathscr{B}} \left[\sum_{k=1}^{\infty} (1 - \lambda_k)^2 \beta_k^2 + J_1(n, \beta) + J_2(n, \beta) + I_2(n, \beta) \right] \\ &\leqslant \sup_{\beta \in \mathscr{B}} \sum_{k=1}^{\infty} \left[(1 - \lambda_k)^2 \beta_k^2 + \frac{\sigma^2}{n} \lambda_k^2 \right] \\ &\quad + \sup_{\beta \in \mathscr{B}} \left[J_1(n, \beta) - \frac{\sigma^2}{n} \sum_{k=1}^{d} \lambda_k^2 + J_2(n, \beta) \right] + \sup_{\beta \in \mathscr{B}} [I_2(n, \beta)]. \end{aligned} \tag{33}$$

The first term on the RHS of (33) is exactly $r_n(\mathscr{B})$ (see [17] or [1]); so in order to complete the proof of the theorem it is sufficient to show that the second and the third terms on the RHS of (33) are of the order $o(r_n)$ as $n \to \infty$.

Due to (9), in order to prove that the second term in (33) is of the order $o(r_n)$ as $n \to \infty$, it is sufficient to show that

$$\sup_{\beta \in \mathscr{B}} \frac{1}{n^2} \sum_{k=1}^{d} \lambda_k^2 \beta_k^2 = o(r_n), \quad n \to \infty, \tag{34}$$

$$\sup_{\beta \in \mathscr{B}} \frac{1}{n} \sum_{k=1}^{d} \lambda_k^2 \sum_{j=d+1}^{\infty} \beta_j^2 = o(r_n), \quad n \to \infty. \tag{35}$$

The proof of (34) is straightforward. Further,

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{d}\lambda_k^2\sum_{j=d+1}^{\infty}\beta_j^2 &\leqslant \frac{L^2}{n}\sum_{k=1}^{d}\lambda_k^2\max_{j>d}[a_j^{-2}] \leqslant \frac{L^2}{n}\sum_{k=1}^{d}\lambda_k^2 v_n^2\\ &\leqslant \frac{1}{n}\sum_{k=1}^{\infty}(1-v_n a_k)_+^2\,\frac{\sigma^2}{n}\sum_{k=1}^{\infty}v_n a_k(1-v_n a_k)_+\\ &\leqslant \frac{1}{n}\sum_{k=1}^{\infty}(1-v_n a_k)_+^2\,\frac{\sigma^2}{n}\sum_{k=1}^{\infty}(1-v_n a_k)_+ \leqslant r_n\frac{d_n}{n}=o(r_n),\\ &\quad n\to\infty\end{aligned}$$

(here we have used the fact that $a_j^{-2}\leqslant v_n^2$ for every $j>d_n$ (by definition of d_n), (5), the fact that $0\leqslant a_k v_n\leqslant 1$ for $k\leqslant d_n$, and (9)). Thus,

$$\sup_{\beta\in\mathscr{B}}\left[J_1(n,\beta)-\frac{\sigma^2}{n}\sum_{k=1}^{d}\lambda_k^2+J_2(n,\beta)\right]=o(r_n),\quad n\to\infty. \tag{36}$$

Now we proceed with bounding $\sup_{\beta\in\mathscr{B}}[I_2(n,\beta)]$. First, by the Cauchy–Schwarz inequality

$$\frac{1}{2}I_2(n,\beta)\leqslant\left(\sum_{k=1}^{d}(1-\lambda_k)^2\beta_k^2\right)^{1/2}\left(\sum_{k=1}^{d}\lambda_k^2[E(\tilde{\beta}_k-\beta_k)]^2\right)^{1/2}\equiv I_{21}(n,\beta)I_{22}(n,\beta). \tag{37}$$

Arguing as before we obtain

$$\begin{aligned}\sup_{\beta\in\mathscr{B}}I_{21}(n,\beta)&=\sup_{\beta\in\mathscr{B}}\left(\sum_{k=1}^{d}(1-\lambda_k)^2\beta_k^2\right)^{1/2}\leqslant\sup_{\beta\in\mathscr{B}}\left(\sum_{k=1}^{\infty}(v_n a_k)^2\beta_k^2\right)^{2}\\ &\leqslant(L^2v_n^2)^{1/2}=\left(v_n\frac{\sigma^2}{n}\sum_{k=1}^{\infty}a_k(1-v_n a_k)_+\right)^{1/2}\leqslant r_n^{1/2}.\end{aligned} \tag{38}$$

Observe that $I_{22}(n,\beta)=||E\tilde{b}_d-b_d||_A$ and our current goal is to bound this quantity from above. Let $\mathscr{F}_{\mathrm{x}}^n$ denote the σ-field generated by n independent sequences $\{x_1(t),x_2(t),\ldots\}$, $t=1,\ldots,n$. Since ϵ is independent of $\{x_k\}_{k=1,2,\ldots}$ we have from (18)

$$E[E(\tilde{b}_d-b_d|\mathscr{F}_{\mathrm{x}}^n)]=E\left[Q_d^{-1}(n)\left(-n^{-1}b_d+\frac{1}{n}\sum_{t=1}^{n}\phi_d(t)\sum_{k=d+1}^{\infty}\beta_k x_k(t)\right)\right].$$

Hence, by the Jensen inequality,

$$I_{22}(n,\beta) = \left\| E\left[Q_d^{-1}(n)\left(-n^{-1}b_d + \frac{1}{n}\sum_{t=1}^{n} \phi_d(t) \sum_{k=d+1}^{\infty} \beta_k x_k(t) \right)\right]\right\|_\Lambda$$
$$\leqslant E\left\| Q_d^{-1}(n)\left(-n^{-1}b_d + \frac{1}{n}\sum_{t=1}^{n} \phi_d(t) \sum_{k=d+1}^{\infty} \beta_k x_k(t) \right)\right\|_\Lambda .$$

Further, using the same reasoning as in bounding $I_1(n,\beta)$ (see (30)–(32)) we finally obtain

$$I_{22}^2(n,\beta) \leqslant \left[J_1(n,\beta) - \left(1 + \frac{\rho_n^*}{1-\rho_n^*}\right)\frac{\sigma^2}{n}\sum_{k=1}^{d}\lambda_k^2 \right] + J_2(n,\beta).$$

Now taking into account (36) we conclude that $\sup_{\beta\in\mathscr{B}} I_{22}(n,\beta) = o(r_n^{1/2}),\ n\to\infty$, and this along with (37) and (38) implies that

$$\sup_{\beta\in\mathscr{B}} I_2(n,\beta) = o(r_n), \quad n\to\infty . \tag{39}$$

Combining (39), (36), and (33) we complete the proof. □

5.2. Proof of Theorem 2

It is sufficient to consider the predictors $\hat{y}(n+1)$ such that $E|\hat{y}(n+1)|^2 < \infty$, because otherwise the lower bound is obvious. First we note that, for any such predictor $\hat{y} = \hat{y}(n+1)$,

$$E[\hat{y}(n+1) - y(n+1)]^2 = E\left[\hat{y}(n+1) - \sum_{k=1}^{\infty}\beta_k x_k(n+1) - \epsilon(n+1)\right]^2$$
$$= \sigma^2 + E\left[\hat{y}(n+1) - \sum_{k=1}^{\infty}\beta_k x_k(n+1)\right]^2 .$$

Further, $\hat{y}(n+1)$ can be decomposed into a sum of two random variables $\hat{y}'(n+1)$ and $\hat{y}''(n+1)$ such that $\hat{y}'(n+1)$ is the orthogonal projection of $\hat{y}(n+1)$ on $\overline{\text{span}}\{x_1(n+1), x_2(n+1), \dots\}$ for fixed $\mathscr{U}_n$, and $\hat{y}''(n+1)$ is orthogonal to $\overline{\text{span}}\{x_1(n+1), x_2(n+1), \dots\}$ for fixed $\mathscr{U}_n$. Note that $\hat{y}'(n+1)$ has the form

$$\hat{y}'(n+1) = \sum_{k=1}^{\infty} \hat{\beta}_k(\mathscr{U}_n) x_k(n+1),$$

where $\hat{\beta}_k(\mathscr{U}_n)$ are random variables measurable w.r.t. $\mathscr{U}_n$. Therefore,

$$\mathscr{R}[\hat{y};\mathscr{B}] \geqslant \sup_{\beta\in\mathscr{B}} E\left[\hat{y}'(n+1) - \sum_{k=1}^{\infty}\beta_k x_k(n+1)\right]^2$$
$$= \sup_{\beta\in\mathscr{B}} E\|\hat{\beta}-\beta\|_2^2 \geqslant \sup_{\beta\in\mathscr{B}'} E\sum_{k=1}^{d}(\hat{\beta}_k - \beta_k)^2$$

for some sequence $\hat{\beta}\in\ell_2$ measurable w.r.t. $\mathscr{U}_n$ and $\mathscr{B}'=\{\beta\in\mathscr{B}\colon\ \beta_k=0,\ k>d\}$. Thus, it is sufficient to establish a lower bound on $\sup_{\beta\in\mathscr{B}'} E\sum_{k=1}^{d}(\hat{\beta}_k-\beta_k)^2$. The further proof is similar to the proof of Theorem 1 in [1]. The difference is that we have random, non-deterministic regressors, and therefore some modifications are needed in calculations of the expected values. We indicate here these modifications.

The proof is based on bounding the minimax risk from below by the Bayes risk and using the van Trees inequality. Assuming that β_k is a random variable with density μ_k and applying the van Trees inequality we get

$$E(\hat{\beta}_k-\beta_k)^2\geqslant\frac{1}{E[I(\beta_k)]+\mathscr{I}(\mu_k)},$$

where the expectation is taken with respect to the joint distribution of $\mathscr{U}_n,\beta_k$. Here $I(\beta_k)$ is the Fisher information about β_k contained in the observations $\mathscr{U}_n$, and $\mathscr{I}(\mu_k)$ is the Fisher information corresponding to the density μ_k. If (12) is fulfilled, then $\mu_k,\ k=1,\dots,d$, are chosen as $\mu_k(x)=(1/m_k)\mu_0(x/m_k),\ k=1,\dots,d$, where μ_0 is a probability density supported on $[-1,1]$, and $\sum_{k=1}^{d}a_k^2m_k^2\leqslant L^2$. We have

$$\begin{aligned}E[I(\beta_k)]&=\int E_{x,y}\left[\sum_{t=1}^{n}\frac{\partial\log\varphi(y(t)-\sum_{j=1}^{d}\beta_jx_j(t))}{\partial\beta_k}\right]^2\mu_k(\beta_k)\,d\beta_k\\&=\frac{1}{\sigma^4}E\left[\sum_{t=1}^{n}\epsilon(t)x_k(t)\right]^2=\sigma^{-2}n,\end{aligned}\tag{40}$$

where $\varphi(\cdot)$ is the standard normal density. This expression is the same as in the case of the deterministic orthonormal design. Note that $\mathscr{I}(\mu_k)=m_k^{-2}I_0$, where I_0 is the Fisher information corresponding to the density μ_0. Therefore we have, for any prediction method $\hat{y}=\hat{y}(n+1)$,

$$\mathscr{R}[\hat{y};\mathscr{B}]\geqslant\frac{\sigma^2}{n}\sum_{k=1}^{d}\frac{m_k^2I_0^{-1}}{m_k^2I_0^{-1}+\sigma^2n^{-1}}.$$

Choosing $m_k^2=\sigma^2(1-v_na_k)_+(nv_na_k)^{-1}$, we see that $\sum_{k=1}^{d}a_k^2m_k^2=L^2$, and thus under condition (12) we get the desired result

$$\mathscr{R}_n[\hat{y};\mathscr{B}]\geqslant\frac{\sigma^2d_n}{n}(1+o(1))=r_n(1+o(1)),\quad n\to\infty.$$

If condition (11) holds, then the prior distributions μ_k are chosen so that

$$\int x\mu_k(x)\,dx=m_k^2(1-\delta/2),\quad \mathscr{I}(\mu_k)\leqslant m_k^{-2}(1+\delta)$$

for some $\delta\in(0,1)$, and $m=(m_1,\dots,m_d)$ satisfying $\sum_{k=1}^{d}a_k^2m_k^2\leqslant L^2$. Proceeding as in [1, pp. 117–118], and computing the expected value of the Fisher information $E[I(\beta_k)]$ as in (40), we obtain the announced result under condition (11). □

5.3. Proof of Theorem 3

The proof goes along the same lines as the proof of Theorem 1. We omit the proof, outlining the main differences from the proof of Theorem 1.

The main difference is that now the regressor vectors $\phi_d(t)$ are dependent for different $t = 1, \dots, n$. However, they are d-dependent; i.e., vectors $\phi_d(t)$ and $\phi_d(s)$ are independent for $|t-s| > d$. Therefore, the exponential inequalities for deviations of $n^{-1}\sum_{t=1}^{n} u(t-k)u(t-j)$, $k,j = 1, \dots, d$ from their expectations can be written down, and the "good" event similar to Ω_α can be defined (see [7, Lemma 1]). Thus, an analog of Lemma 1 can be established. Further, results similar to Lemmas 2 and 3 are easily obtained. In particular, for

$$I_{11} = \frac{1}{n}\sum_{t=1}^{n} \phi_d(t) \sum_{k=d+1}^{\infty} \beta_k u(t-k), \quad I_{12} = \frac{1}{n}\sum_{t=1}^{n} \phi_d(t)\epsilon(t),$$

the same inequalities (25) and (26) hold true. Other details of the proof remain unchanged. □

Acknowledgments

The research was supported in part by a Grant of the ESF programme on "Highly Structured Stochastic Systems (HSSS)" and by a Grant of the Arc-en-ciel/Keshet program.

References

[1] E.N. Belitser, B.Y. Levit, Asymptotically minimax nonparametric regression in L_2, Statistics 28 (1996) 105–122.
[2] L. Breiman, D. Friedman, How many variables should be entered in a regression equation?, J. Amer. Statist. Assoc. 78 (1983) 131–136.
[3] P.J. Brockwell, R.A. Davis, Time Series: Theory and Methods, 2nd Edition, Springer, New York, 1991.
[4] S. Efromovich, On nonparametric regression for iid observations in a general setting, Ann. Statist. 24 (1996) 1126–1144.
[5] S.Y. Efroimovich, M.S. Pinsker, Estimation of square–integrable probability density of a random variable, Problems Inform. Transmission 18 (1982) 175–182.
[6] S.Yu. Efroimovich, M.S. Pinsker, Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression, Statist. Sinica 6 (1996) 925–942.
[7] A. Goldenshluger, Nonparametric estimation of transfer functions: rates of convergence and adaptation, IEEE Trans. Inform. Theory 44 (1998) 644–658.
[8] A. Goldenshluger, A. Tsybakov, Adaptive prediction and estimation in linear regression with infinitely many parameters, Ann. Statist. 29 (2001) 1601–1619.
[9] A. Goldenshluger, A. Zeevi, Non-asymptotic bounds for autoregressive time series modeling, preprint, 1999.
[10] G.K. Golubev, M.S. Pinsker, Minimax extrapolation of sequences, Problems Inform. Transmission 19 (1983) 275–283.

[11] G.K. Golubev, M.S. Pinsker, Extremal properties of minimax estimation of sequences, Problems Inform. Transmission 21 (1985) 192–206.
[12] G.K. Golubev, On minimax filtering of functions in L_2, Problems Inform. Transmission 18 (1982) 67–75.
[13] P. Huber, Robust regression: asymptotics, conjectures and Monte-Carlo, Ann. Statist. 1 (1973) 799–821.
[14] T.L. Lai, C.Z. Wei, Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems, Ann. Statist. 10 (1982) 154–166.
[15] M. Nussbaum, Spline smoothing in regression models and asymptotic efficiency in L_2, Ann. Statist. 13 (1985) 984–992.
[16] V.V. Petrov, Limit Theorems of Probability Theory, Clarendon, Oxford, 1995.
[17] M.S. Pinsker, Optimal filtering of square integrable signals in Gaussian white noise, Problems Inform. Transmission 16 (1980) 120–133.
[18] S. Portnoy, Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency, Ann. Statist. 12 (1984) 1298–1309.
[19] S. Portnoy, Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. II. Normal approximation, Ann. Statist. 13 (1985) 1403–1417.
[20] R. Shibata, An optimal selection of regression variables, Biometrika 68 (1981) 45–54.
[21] A.B. Tsybakov, Asymptotically efficient estimation of a signal in L_2 under general loss functions, Problems Inform. Transmission 33 (1997) 78–88.
[22] V.J. Yohai, R.A. Maronna, Asymptotic behavior of M-estimators for the linear model, Ann. Statist. 7 (1979) 258–268.